**Lang\_cms自动化采集程序0.92版本使用说明书**

**前言：使用前请先阅读本文档。**

**本程序基于Python + Mysql结合，用于实现CMS全自动采集。**

**此软件用于实现以下功能:**

**1 url无限采集**

**2 CMS 识别**

**3备份文件扫描（包含svn源码泄露扫描）**

**4 扫描使用了ST2框架的网址**

**配置文件**

**cms.txt >>>>>> cms识别文件，可以自己添加新的规则**

**rar.txt >>>>>> 备份文件列表，可以自定义添加**

**Config.ini>>>>> 数据库及软件参数配置文件**

**mysql.sql>>>>> 数据库文件**

**第一步安装环境**

**使用前先安装好Mysql数据库环境**

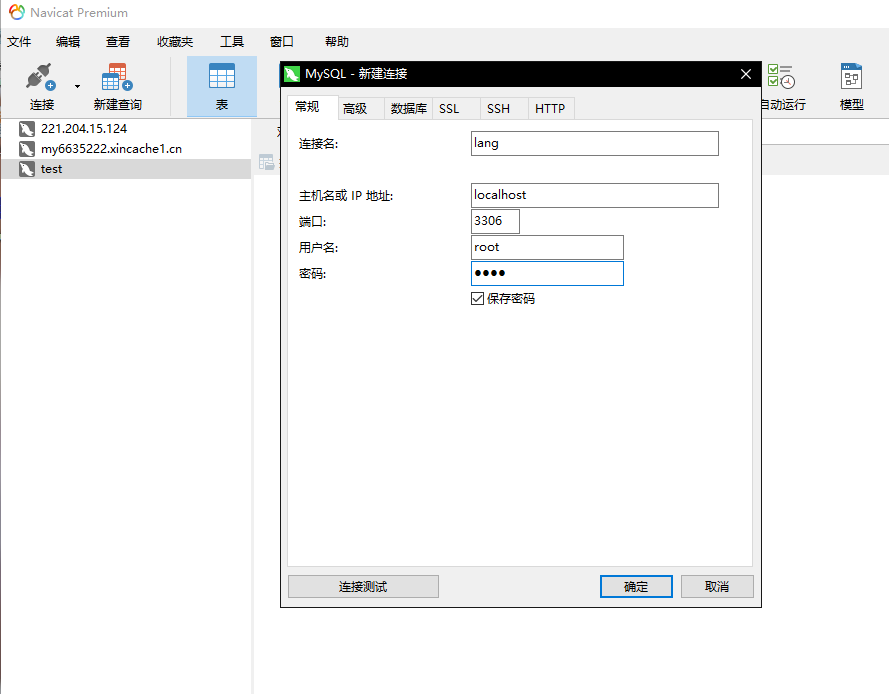
**建议使用phpstduy，搭建起来更加简单友好**

**下载地址：**[**http://www.phpstudy.net/**](http://www.phpstudy.net/)

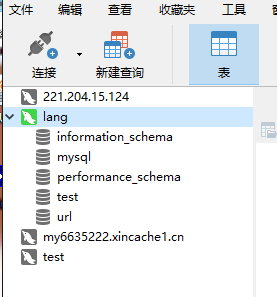
**安装好之后，下载数据库管理软件，个人建议navicat.exe**

**下载地址：**[**https://www.navicat.com/en/products**](https://www.navicat.com/en/products)

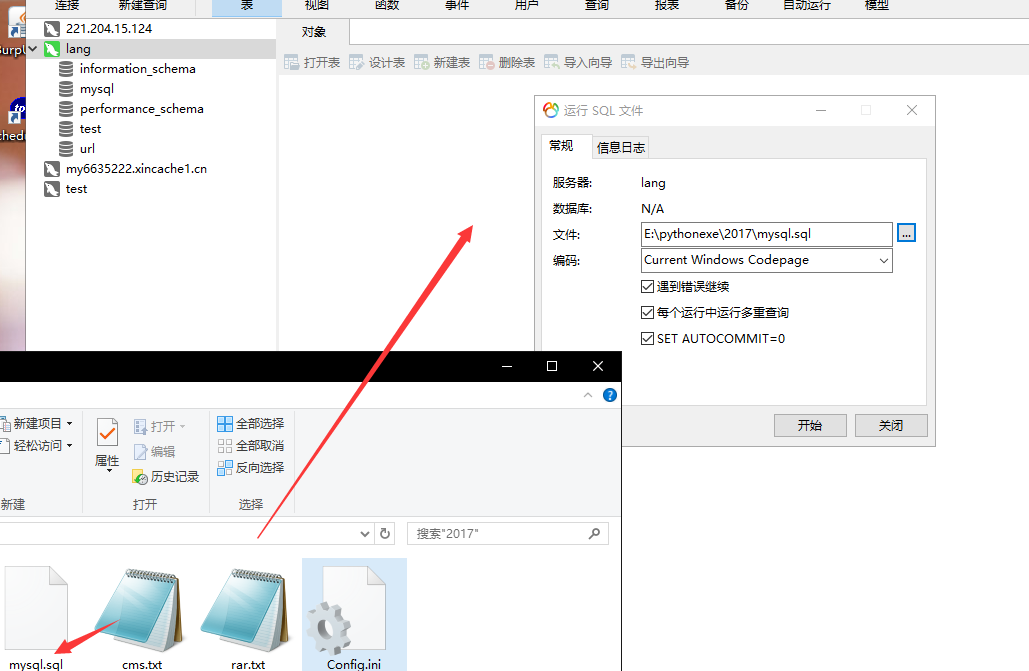
**然后运行phpstudy确保开启MYSQL 打开navicat.exe 默认账号密码root/root**



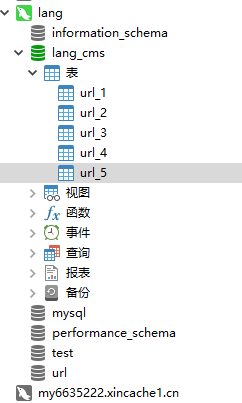
**点击确定连接上去 双击lang 会看到如下界面**



**把mysql.sql直接拖进去，点击开始**



**然后刷新，当出现下这个界面的时候就说明环境都配置完毕**



**到这里需要下载的软件和环境已经全部安装完毕。**

**第二步你需要一些网址作为你的原始爬行目标，你可以用一些url采集工具采集一些网站，比如下图。这些网址都是你的原始网址，无限爬行基于这些网址的基础上爬行，爬行到的网址基本上都是这一类型的，比如你的原始网址都带有EDU,那么无限爬行到底网址大部分都是EDU网站。如果你想要弄一些和别人不一样的网址，你可以自定义关键词或者手动在百度上找一些网址。**



**然后导出关键词，把导出的关键词文本复制到主程序目录下。然后配置好相关文件Config.ini**

**Config.ini 详细使用说明：**

**[Server]**

**Host=127.0.0.1 #数据库地址 如果无法连接就改成localhost**

**Username=root #数据库账号密码**

**Password=root**

**db=Lang\_cms**

**[Config]**

**New\_start=1 #重置自定义外部链接 1是自定义 0是不接受自定义外部链接在数据库无限采集 第一次使用请设置为1 如果想添加新的外部链接 改成0然后重启主程序，简而言之，就是你第一次使用就设置1，这样就会让你输入自己采集好的url文本，运行过程中会自动改成0，改成0后再次点击程序就不会让你输入采集好的url文本，这样做的目的就是当你重启电脑后，第二次运行直接双击就可以了。**

**Rarscan=1 #扫描备份文件和svn源码泄露 1是开启 0是关闭**

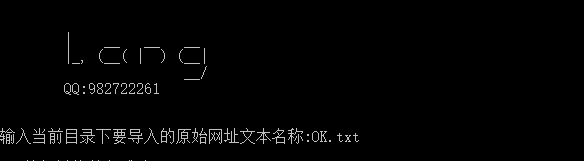
**St2scan=1 #扫描使用st2框架搭建的网址 1是开启 0是关闭**

**Cmsscan=1 #扫描Cms类型 1是开启 0是关闭**

**urlcaiji = 1 #自动url采集 1是开启 0是关闭**

**thread\_s = 10 #线程数 一般设置10就差不多够了**

**第三步你直接运行主程序就可以了，双击打开主程序，出现如下界面。**



**第一次运行的时候会让你输入名称，这里输入第二步采集好了的文本名称就可以了，然后回车。这里就开始对原始网站开始爬行了。如果你的原始网站比较多的话，这里可能会爬行得比较久。**



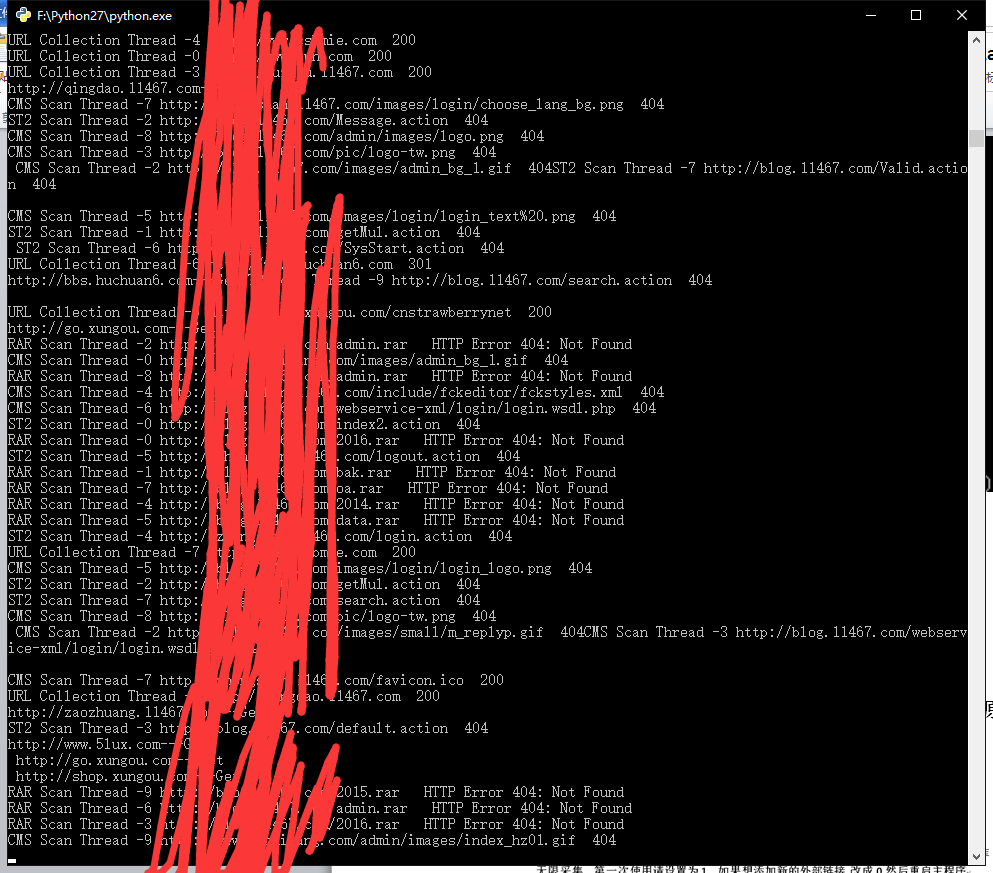
**到这里所有的原始网址爬行完毕，对爬行结果进行存活检测，然后存储到数据库**



**存活检测**



**然后剩下的就不用管了**



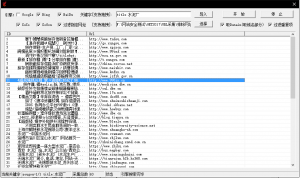
**常见问题：**

1 我不会搭建mysql数据库环境和库怎么办？

这个好办，把mysql搭建过程写在帮助文档就好了。

2 我不想和别人采集到一样的网站怎么办？

我的解决办法是：自定义初始链接。**你可以用一些url采集工具采集一些网站，比如下图。这些网址都是你的原始网址，无限爬行基于这些网址的基础上爬行，爬行到的网址基本上都是这一类型的，比如你的原始网址都带有EDU,那么无限爬行到底网址大部分都是EDU网站。如果你想要弄一些和别人不一样的网址，你可以自定义关键词或者手动在百度上找一些网址然后保存在一个文本里面。**



3 我想在数据库查询数据怎么办？

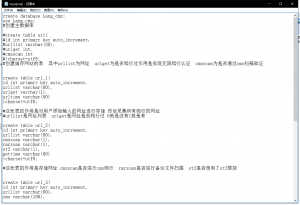
打开数据库管理软件，然后再Lang\_cms这个数据库里面查看，但是有一些朋友对mysql并不是很熟悉。于是我想到帮写一个专门查询mysql数据库内容的软件，包含以下功能。当然这只是一些粗略的功能，至于查询使用了什么cms  大家可以这样查询 select \* from url\_3 whrer cms = "dedecms";  可以把dedecms改成别的



4 数据库结构？

这个的确让我花费了很多心思，我觉得写软件构思和编程能力同样重要。最近在学习python算法，里面讲得编程的那种思维对我影响很大，如何更快更稳的运行你的程序。

我觉得创建5个表，分别存储不同的数据。把这个sql直接拖进去安装就好了。



5 这种无限采集的程序比较麻烦，第一次运行和第二次运行有什么区别吗？

这么说吧....第一次运行的时候，需要你提供原始url文本，在这些文本的基础上爬行。

第二次运行的时候（因为你已经运行过一次了，废话），这个时候数据库已经保存了很多爬行的网址，就不需要你在载入外部链接了，程序直接运行。

6  如果我想把这一切数据都删除推到重新扫描怎么办？

当然是先删除数据库啊，如果你不会.....就打开sql查询帮手就，看到7或8没，你可以选择删除，7的效果是不删除数据库，只是把表里面所有的内容一次性全部清空，

8的效果是直接删除数据库，这下数据库就没了，你需要重新把sql文件拖进去执行一遍。

7  如果我第二次运行想继续添加自己采集到的外部url文本怎么办？

这个好吧，我先给你介绍一下配置文件Config.ini

**Config.ini  详细使用说明：**

**[Server]**

**Host=127.0.0.1   #数据库地址 如果无法连接就改成localhost**

**Username=root    #数据库账号密码**

**Password=root**

**db=Lang\_cms**

**[Config]**

**New\_start=1   #重置自定义外部链接 1是自定义  0是不接受自定义外部链接在数据库无限采集  第一次使用请设置为1  如果想添加新的外部链接 改成0然后重启主程序，简而言之，就是你第一次使用就设置1，这样就会让你输入自己采集好的url文本，运行过程中会自动改成0，改成0后再次点击程序就不会让你输入采集好的url文本，这样做的目的就是当你重启电脑后，第二次运行直接双击就可以了。**

**Rarscan=1        #扫描备份文件和svn源码泄露  1是开启 0是关闭**

**St2scan=1        #扫描使用st2框架搭建的网址  1是开启 0是关闭**

**Cmsscan=1      #扫描Cms类型  1是开启 0是关闭**

**urlcaiji = 1        #自动url采集  1是开启 0是关闭**

**thread\_s = 10  #线程数  一般设置10就差不多够了**

看到**New\_start 不解释太多**

8  我的程序运行很慢怎么办？

软件耗时基本都花费在网络连接上，有可能是你的原始网址里面的链接有防护阻止访问或者链接已经死了没办法访问。换新的网址就好了或者换宽带好了。

9  我的程序不动了怎么办？

回车键敲一下

一些常见问题大概就这么多。

**当前版本为0.92测试版，大家发现BUG或者有什么想要添加的功能请联系浪子本人**

**QQ：982722261**